BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59167
doi: 10.3897/biss.4.59167

OPEN ACCESS

Conference Abstract

Slinging With Four Giants on a Quest to Credit Natural Historians for our Museums and Collections

David Peter Shorthouse ‡

‡ Agriculture & Agri-Food Canada, Ottawa, Canada

Corresponding author: David Peter Shorthouse (davidpshorthouse@gmail.com)

Received: 30 Sep 2020 | Published: 02 Oct 2020

Citation: Shorthouse DP (2020) Slinging With Four Giants on a Quest to Credit Natural Historians for our Museums and Collections. Biodiversity Information Science and Standards 4: e59167.
https://doi.org/10.3897/biss.4.59167

Abstract

Bionomia, https://bionomia.net previously called Bloodhound Tracker, was launched in August 2018 with the aim of illustrating the breadth and depth of expertise required to collect and identify natural history specimens represented in the Global Biodiversity Information Facility (GBIF). This required that specimens and people be uniquely identified and that a granular expression of actions (e.g. "collected", "identified") be adopted. The Darwin Core standard presently combines agents and their actions into the conflated terms recordedBy and identifiedBy whose values are typically unresolved and unlinked text strings. Bionomia consists of tools, web services, and a responsive website, which are all used to efficiently guide users to resolve and unequivocally link people to specimens via first-class actions collected or identified. It also shields users from the complexity of casting together and seamlessly integrating the services of four giant initiatives: ORCID, Wikidata, GBIF, and Zenodo. All of these initiatives are financially sustainable and well-used by many stakeholders, well-outside this narrow user-case. As a result, the links between person and specimen made by users of Bionomia are given every opportunity to persist, to represent credit for effort, and to flow into collection management systems as meaningful new entries. To date, 13M links between people and specimens have been made including 2M negative associations on 12.5M specimen records. These links were either made by the

This is an open access article distributed under the terms of the CC0 Public Domain Dedication.

collectors themselves or by 84 people who have attributed specimen records to their peers, mentors and others they revere.

Integration With ORCID and Wikidata

People are identified in Bionomia through synchronization with ORCID and Wikidata by reusing their unique identifiers and drawing in their metadata. ORCID identifiers are used by living researchers to link their identites to their research outputs. ORCID services include OAuth2 pass-through authentication for use by developers and web services for programmatic access to its store of public profiles. These contain elements of metadata such as full name, aliases, keywords, countries, education, employment history, affiliations, and links to publications. Bionomia seeds its search directory of people by periodically querying ORCID for specific user-assigned keywords as well as directly though account creation via OAuth2 authentication. Deceased people are uniquely identified in Bionomia through integration with Wikidata by caching unique 'Q' numbers (identifiers), full names and aliases, countries, occupations, as well as birth and death dates. Profiles are seeded from Wikidata through daily queries for properties that are likely to be assigned to collectors of natural history specimens such as "Entomologists of the World ID" (= P5370) or "Harvard Index of Botanists ID" (= P6264). Because Wikidata items may be merged, Bionomia captures these merge events, re-associates previously made links to specimen records, and mirrors Wikidata's redirect behaviour. A Wikidata property called "Bionomia ID" (= P6944), whose values are either ORCID identifiers or Wikidata 'Q' numbers, helps facilitate additional integration and reuse.

Integration with GBIF

Specimen data are downloaded wholesale as Darwin Core Archives from GBIF every two weeks. The purpose of this schedule is to maintain a reasonable synchrony with source data that balances computation time with the expections of users who desire the most up-to-date view of their specimen records. Collectors with ORCID accounts who have elected to receive notice, are informed via email message when the authors of newly published papers have made use of their specimen records downloaded from GBIF.

Integration with Zenodo

Finally, users of Bionomia may integrate their ORCID OAuth2 authentication with Zenodo, an industry-recognized archive for research data, which enjoys support from the Conseil Européen pour la Recherche Nucléaire (CERN). At the user's request, their specimen data represented as CSV (comma-separated values) and JSON-LD (JavaScript Object Notation for Linked Data) documents are pushed into Zenodo, a DataCite DOI is assigned, and a formatted citation appears on their Bionomia profile. New versions of these files are pushed to Zenodo on the user's behalf when new specimen records are linked to them. If users have configured their ORCID account to listen for new entries in DataCite, a new work entry will also be made in their ORCID profile, thus sealing a perpetual, semi-automated loop betwen GBIF and ORCID that tidily showcases their efforts at collecting and identifying natural history specimens.

Technologies Used

Bionomia uses Apache Spark via scripts written in Scala, a human name parser written in Ruby called dwc_agent, queues of jobs executed through Sidekiq, scores of pairwise similarities in the structure of human names stored in Neo4j, data persistence in MySQL, and a search layer in Elasticsearch.

Here, I expand on lessons learned in the construction and maintenance of Bionomia, emphasize the criticality of recognizing the early efforts made by a fledgling community of enthusiasts, and describe useful tools and services that may be integrated into collection management systems to help churn strings of unresolved, unlinked collector and determiner names into actionable identifiers that are gateways to rich sources of information.

Keywords

credit, occurrence, ORCID, wikidata, GBIF, Bionomia

Presenting author

David Peter Shorthouse

Presented at

TDWG 2020